

# Best Friends Forever (BFF): Finding Lasting Dense Subgraphs

Konstantinos Semertzidis<sup>1</sup>, Evaggelia Pitoura<sup>1</sup>, Evimaria Terzi<sup>2</sup>, Panayiotis Tsaparas<sup>1</sup>

<sup>1</sup>Computer Science and Engineering Department, University of Ioannina, Greece

<sup>2</sup>Computer Science Department, Boston University, USA

{ksemer, pitoura, tsap}@cs.uoi.gr, evimaria@cs.bu.edu

## ABSTRACT

Graphs form a natural model for relationships and interactions between entities, for example, between people in social and cooperation networks, servers in computer networks, or tags and words in documents and tweets. But, which of these relationships or interactions are the most lasting ones? In this paper, given a set of graph snapshots, which may correspond to the state of a dynamic graph at different time instances, we look at the problem of identifying the set of nodes that are the most densely connected at all snapshots. We call this problem the *Best Friends For Ever* (BFF) problem. We provide definitions for density over multiple graph snapshots, that capture different semantics of connectedness over time, and we study the corresponding variants of the BFF problem. We then look at the *On-Off* BFF ( $O^2$ BFF) problem that relaxes the requirement of nodes being connected in all snapshots, and asks for the densest set of nodes in at least  $k$  of a given set of graph snapshots. We show that this problem is NP-complete for all definitions of density, and we propose a set of efficient algorithms. Finally, we present experiments with synthetic and real datasets that show both the efficiency of our algorithms and the usefulness of the BFF and the  $O^2$ BFF problems.

## 1. INTRODUCTION

Graphs offer a natural model for capturing the interactions and relationships among entities. Oftentimes, multiple snapshots of a graph are available; for example, these snapshots may correspond to the states of a dynamic graph at different time instances. We call such sets of graph snapshots, a *graph history*. Analysis of the graph history finds a large spectrum of applications, ranging from social-network marketing, to virus propagation and digital forensics. A central question in this context is: *which interactions, or relationships in a graph history are the most lasting ones?*

In this paper, we formalize this question and we design algorithms that effectively identify such relationships.

In particular, given a graph history, we introduce the problem of efficiently finding the set of nodes, that remains the most tightly connected through history. We call this problem the *Best Friends For Ever* (BFF) problem. We formulate the BFF problem as the problem of locating the set of nodes that have the maximum aggregate density in the graph history. We provide different definitions for the aggregate density that capture different notions of connectedness over time, and result to four variants of the BFF problem. For two of them, we show that they can be solved optimally in polynomial time. For the other two, we propose efficient greedy algorithms. We then extend the problem so as to locate the BFFs of a specific set of input query nodes.

Identifying BFF nodes finds many applications. For example, in collaboration and social networks such nodes can be chosen to organize successful professional or social events; usually the success of such events depends on whether the participants are well-acquainted with each other. In a protein-protein network, we can locate protein complexes that are densely interacting at different states, thus indicating a possible underlying regulatory mechanism. In a network where nodes are words or tags and edges correspond to their co-occurrences in documents or tweets published in a specific period of time, identifying BFF may serve as a first step in topic identification, tag recommendation and other types of analysis. In a computer network, locating servers that communicate heavily over time may be useful in identifying potential attacks, or bottlenecks.

We extend the BFF problem to capture the cases where subsets of nodes are densely connected for only a subset of the snapshots. Consider for example, a set of collaborators that work intensely together for some years and then they drift apart, or, a set of friends in a social network that stop interacting for a few snapshots and then, they reconnect with each other. To identify such subsets of nodes, we define the *On-Off* BFF problem, or  $O^2$ BFF for short. In the  $O^2$ BFF problem, we relax the requirement of nodes being connected in all snapshots, and ask for the densest set of nodes in at least  $k$  of the snapshots. Depending again on the definition of aggregate density, we get different variants of the  $O^2$ BFF problem. We show that all variants are NP-hard and propose efficient iterative algorithms.

Our experimental results with real and synthetic datasets show the efficiency and effectiveness of our algorithms in discovering lasting dense subgraphs. Two case studies on bibli-

ographic collaboration networks, and hashtag co-occurrence networks in *Twitter* validate our approach.

The problem of identifying a dense subgraph in a static (i.e., single-snapshot) graph has received a lot of attention (e.g., [6, 8, 10]). However, to the best of our knowledge, we are the first to systematically introduce and study density in a graph history and define the BFF and O<sup>2</sup>BFF problems and their variants. The most related work to ours [9] studies a problem related to just one of the four variants of the BFF problem in the context of graph databases.

**Contributions:** To summarize, the main contributions of this work are:

- We introduce the BFF problem of identifying a subset of nodes that define dense subgraphs in all snapshots of a graph history. We extend the notion of density to capture different semantics of density over time that lead to four variants of the BFF problem. We study the complexity of the variants of the BFF problem and propose appropriate linear-time algorithms.
- We define the O<sup>2</sup>BFF problem of identifying both a subset of nodes  $S$  and a subset of snapshots  $T$  such that the nodes in  $S$  define dense subgraphs in the snapshots in  $T$ . We show that for all the aggregate density functions we consider the corresponding O<sup>2</sup>BFF problems are NP-hard and design iterative polynomial-time heuristic algorithms.
- We extend our definitions and algorithms to identify the BFFs of an input set of query nodes.
- Our experiments with both real and synthetic datasets demonstrate that our problem definitions are meaningful and that our algorithms work well in identifying dense subgraphs in practice.

**Roadmap:** The BFF problem is introduced in Section 2 and its algorithms in Section 3. In Section 4, we present the O<sup>2</sup>BFF problem and algorithms for solving it, while in Section 5 we study extensions to the original problem. Our experimental evaluation is presented in Section 6 and comparison with related work in Section 7. Section 8 concludes the paper.

## 2. THE BFF PROBLEM

In this section, we introduce the BFF problem and aggregate density over time.

### 2.1 Problem definition

Graphs are a natural model of relationships and interactions among entities. In this paper, we assume that we are given multiple snapshots of these interactions or relationships. Snapshots may be ordered, for example, when the snapshots correspond to the states of a dynamic graph as the graph evolves over time. We may also have unordered snapshots, e.g., a collection of graphs, for example when the snapshots correspond to graphs collected as a result of some scientific experiments. For notational simplicity, we assume that the graph snapshots are over the same set of nodes  $V$ , but our work is applicable to graph snapshots with different set of nodes by considering  $V$  as their union.

**Definition 1 (GRAPH HISTORY).** A graph history  $\mathcal{G} = \{G_1, G_2, \dots, G_\tau\}$  over  $\tau$  instances is a collection of graph snapshots, where each snapshot  $G_t = (V, E_t)$ ,  $t \in [1, \tau]$ , is defined over the same set of nodes  $V$ .

An example of a graph history with four snapshots is shown in Figure 1.

Given the snapshots of a graph history  $\mathcal{G}$ , our goal is to locate the Best Friends For Ever (BFF) that is to identify a subset of nodes of  $V$  such that these nodes remain densely connected with each other in *all* snapshots of  $\mathcal{G}$ . Formally:

**Problem 1 (The Best Friends Forever (BFF) Problem).** Given a graph history  $\mathcal{G}$  and an aggregate density function  $f$ , find a subset of nodes  $S \subseteq V$ , such that  $f(S, \mathcal{G})$  is maximized.

### 2.2 Aggregate density

We start by reviewing two basic definitions of graph density. Given an undirected graph  $G = (V, E)$  and a node  $u$  in  $V$ , we use  $\text{degree}(u, G)$  to denote the degree of  $u$  in  $G$ .

The *average density* of the graph  $d_a(G)$  is the average degree of the nodes in  $V$ :

$$d_a(G) = \frac{1}{|V|} \sum_{u \in V} \text{degree}(u, G) = \frac{2|E|}{|V|},$$

while the *minimum density* of the graph  $d_m(G)$  is the minimum degree of any node in  $V$ :

$$d_m(G) = \min_{u \in V} \text{degree}(u, G).$$

Intuitively, for a given graph,  $d_m$  is defined by a single node, the one with the minimum degree, while  $d_a$  accounts for the degrees and thus the connectivity of all nodes. For example, in Figure (1a)  $d_m(G_1) = 2$ , while  $d_a(G_1) = 10/3$ . Clearly,  $d_m$  is a lower bound for  $d_a$ . From now on, when the subscript of  $d$  is ignored, density can be either  $d_a$  or  $d_m$ .

We also define the density of a subset of nodes  $S \subseteq V$  in the graph  $G = (V, E)$ . To this end, we use the induced subgraph  $G[S] = (S, E(S))$  in  $G$ , where  $E(S) = \{(u, v) \in E : u \in S, v \in S\}$ . We define the density  $d(S, G)$  of  $S$  in  $G$  as  $d(G[S])$ . For example, again for snapshot  $G_1$  in Figure 1, for  $S_x = \{x_1, x_2, x_3, x_4\}$ ,  $d_m(S_x, G_1) = d_a(S_x, G_1) = 3$ , while for  $S_y = \{y_1, y_2, y_3, y_4, y_5\}$ ,  $d_m(S_y, G_1) = 2$  and  $d_a(S_y, G_1) = 16/5$ . Between  $S_x$  and  $S_y$ ,  $S_x$  has the highest minimum density, whereas  $S_y$  the highest average density.

We now define the density of a set of nodes  $S$  on a graph history. To do this, we need a way to aggregate the density of a set of nodes over multiple graph snapshots.

**Aggregating density sequences:** Given a graph history  $\mathcal{G} = \{G_1, \dots, G_\tau\}$ , we will use  $d(S, \mathcal{G}) = \{d(S, G_1), \dots, d(S, G_\tau)\}$  to denote the sequence of density values for the graph induced by the set  $S$  in the graph snapshots. We consider two definitions for an *aggregation function*  $g(d(S, \mathcal{G}))$  that aggregates the densities over snapshots: the first,  $g_m$ , computes the minimum density over all snapshots:

$$g_m(d(S, \mathcal{G})) = \min_{G_t \in \mathcal{G}} d(S, G_t),$$

while the second,  $g_a$ , computes the average density over all snapshots:

$$g_a(d(S, \mathcal{G})) = \frac{1}{|\mathcal{G}|} \sum_{G_t \in \mathcal{G}} d(S, G_t).$$

Intuitively, the minimum aggregation function requires high density in each and every snapshot, while the average aggregation function looks at the snapshots as a whole. Again, we use  $g$  to collectively refer to  $g_m$  or  $g_a$ . We can now define the *aggregate density*  $f$ .

**Definition 2 (AGGREGATE DENSITY).** Given a graph history  $\mathcal{G} = \{G_1, \dots, G_\tau\}$  defined over a set of nodes  $V$  and  $S \subseteq V$ , we define the *aggregate density*  $f(S, \mathcal{G})$  to be

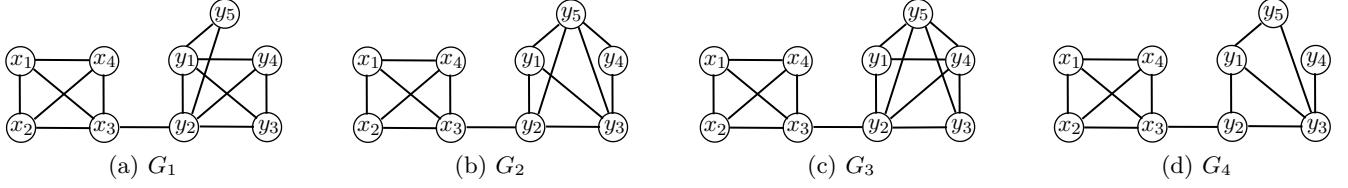


Figure 1: A graph history  $\mathcal{G} = \{G_1, \dots, G_4\}$  consisting of four snapshots.

$f(S, \mathcal{G}) = g(d(S, \mathcal{G}))$ . Depending on the choice of the density function  $d$  and the aggregation function  $g$ , we have the following four versions of  $f$ : (a)  $f_{mm}(S, \mathcal{G}) = g_m(d_m(S, \mathcal{G}))$ , (b)  $f_{ma}(S, \mathcal{G}) = g_m(d_a(S, \mathcal{G}))$ , (c)  $f_{am}(S, \mathcal{G}) = g_a(d_m(S, \mathcal{G}))$ , and (d)  $f_{aa}(S, \mathcal{G}) = g_a(d_a(S, \mathcal{G}))$ .

By considering the four choices for the aggregate density function  $f$ , we have four variants of the BFF problem. Specifically,  $f_{mm}$ ,  $f_{ma}$ ,  $f_{am}$  and  $f_{aa}$  give rise to problems: BFF-MM, BFF-MA, BFF-AM and BFF-AA respectively. Each of them associates different semantics with the meaning of density among nodes in a graph history.

Large values of  $f_{mm}(S, \mathcal{G})$  correspond to groups of nodes  $S$  where each member of the group is connected with a large number of other members of the group at each snapshot. A node ceases to be considered a member of the group, if it loses touch with the other members even in a single snapshot. We expect such groups of friends to be small in size.

Large values of  $f_{ma}(S, \mathcal{G})$  are achieved for groups with high average density at each snapshot  $G \in \mathcal{G}$ . As opposed to  $f_{mm}(S, \mathcal{G})$ , where the requirement is placed at each member of the group, large values of  $f_{ma}(S, \mathcal{G})$  are indicative that the group  $S$  has persistently high density as a whole.

The  $f_{aa}(S, \mathcal{G})$  metric takes large values when the group  $S$  has many connections on average; thus,  $f_{aa}$  is more “loose” both in terms of consistency over time and in terms of requirements at the individual group member level.

For example, in the graph history  $\mathcal{G}$  in Figure 1, all aggregate densities for  $S_x$  are equal to 3. However, for  $S_y$   $f_{aa}(S_y, \mathcal{G}) = 31/10$ , while  $f_{ma}(S_y, \mathcal{G}) = 12/5$ . That is, while  $S_y$  has still larger average density than  $S_x$  if we consider  $f_{aa}$ ,  $S_x$  has now larger average density in terms of  $f_{ma}$ , due to the last instance. Note also that  $f_{mm}(S_y, \mathcal{G}) = 1$  and that this value is determined by just one node in just one snapshot, i.e., node  $y_4$  in the last snapshot.

Lastly,  $f_{am}(S, \mathcal{G})$  takes the average of the minimum degree node at each snapshot, thus is less sensitive to the density of  $S$  at a single instance. In Figure 1,  $f_{am}(S_y, \mathcal{G}) = 2$ .

**The average graph:** Finally, let us introduce the *average graph* of a graph history  $\mathcal{G}$  which is an edge-weighted graph where the weight of an edge is equal to the fraction of snapshots in  $\mathcal{G}$  where the edge appears.

**Definition 3 (AVERAGE GRAPH).** Given a graph history  $\mathcal{G} = \{G_1, \dots, G_\tau\}$  on a set of nodes  $V$ , the average graph  $\hat{H}_{\mathcal{G}} = (V, \hat{E}, \hat{w})$  is a *weighted, undirected* graph on the set of nodes  $V$ , where  $\hat{E} = V \times V$ , and for each  $(u, v) \in \hat{E}$ ,  $\hat{w}(u, v) = \frac{|G_t = (V, E_t) \in \mathcal{G} | (u, v) \in E_t|}{|\mathcal{G}|}$ .

As usual, the degree of a node  $u$  in a weighted graph is defined as:  $\text{degree}(u, \hat{H}_{\mathcal{G}}) = \sum_{(u, v) \in \hat{E}} \hat{w}(u, v)$ . Note that the average graph performs aggregation on a per-node basis, in

that, the degree of each node  $u$  in  $\hat{H}_{\mathcal{G}}$  is the average degree of  $u$  in time. With some algebraic manipulation, we can show that:

**Lemma 1.** Let  $\mathcal{G} = \{G_1, \dots, G_\tau\}$  be a graph history over a set of nodes  $V$  and  $S$  a subset of nodes in  $V$ , it holds:  $f_{aa}(S, \mathcal{G}) = d_a(\hat{H}_{\mathcal{G}}[S])$ .

### 3. BFF ALGORITHMS

We now introduce a generic algorithm for the BFF problem. The algorithm (shown in Algorithm 1) starts with a set of nodes  $S_0$  consisting of all nodes  $V$ , and then it performs  $n - 1$  steps, where at each step  $i$  it produces a set  $S_i$  by removing one of the nodes in the set  $S_{i-1}$ . It then finds the set  $S_i$  with the maximum aggregate density  $f(S, \mathcal{G})$ .

---

**Algorithm 1** The FINDBFF algorithm.

---

**Input:** Graph history  $\mathcal{G} = \{G_1, \dots, G_\tau\}$ ; an aggregate density function  $f$

**Output:** A subset of nodes  $S$

---

```

1:  $S_0 = V$ 
2: for  $i = 1, \dots, n - 1$  do
3:    $v_i = \text{SELECTNODE}(S_{i-1})$ 
4:    $S_i = S_{i-1} \setminus \{v_i\}$ 
5: end for
6: return  $\arg \max_{i=0 \dots n-1} f(S_i, \mathcal{G})$ 
```

---

The FINDBFF algorithm forms the basis for all the algorithms that we propose for the four variants of the BFF problem. We get different instantiations of the FINDBFF algorithm depending on the criterion based on which the SELECTNODE procedure in line 3 decides which node to remove at each step. In the following we introduce different criteria for the SELECTNODE procedure appropriate for each of the four variants of the BFF problem.

#### 3.1 Solving BFF-MM

For each node  $v$  in  $S$ , we define its  $\text{score}_m$  in graph history  $\mathcal{G}$  to be equal to the minimum degree of  $v$ ,

$$\text{score}_m(v, \mathcal{G}[S]) = \min_{G_t \in \mathcal{G}} \text{degree}(v, G_t[S]).$$

Then at the  $i$ -th iteration of the FINDBFF algorithm, we select the node  $v_i$  such that

$$v_i = \arg \min_{v \in S_{i-1}} \text{score}_m(v, \mathcal{G}[S_{i-1}]).$$

We call this instantiation of the FINDBFF algorithm FINDBFF<sub>MM</sub>. Below we prove that FINDBFF<sub>MM</sub> provides the optimal solution to the BFF-MM problem.

**Proposition 1.** The BFF-MM problem can be solved optimally in polynomial time using the FINDBFF<sub>MM</sub> algorithm.

*Proof.* Let  $i$  be the iteration of the  $\text{FINDBFF}_M$  algorithm, where for the first time, a node that belongs to an optimal solution  $S^*$  is selected to be removed. Let  $v_i$  be this node. Then clearly,  $S^* \subseteq S_{i-1}$  and by the fact that at every iteration we remove edges from the graphs we have that

$$\text{score}_m(v_i, \mathcal{G}[S_{i-1}]) \geq \text{score}_m(v_i, \mathcal{G}[S^*]).$$

Since  $v_i$  is the node we pick at iteration  $i$ , every node  $u \in S_{i-1}$  satisfies:

$$\min_{G_t \in \mathcal{G}} \text{degree}(u, G_t[S_{i-1}]) = \text{score}_m(u, \mathcal{G}[S_{i-1}]) \geq \text{score}_m(v_i, \mathcal{G}[S_{i-1}]) \geq \text{score}_m(v_i, \mathcal{G}[S^*]).$$

Since this is true for every node  $u$ , this means that  $S_{i-1}$  is indeed optimal and that our algorithm will find it.  $\square$

Note that an algorithm that iteratively removes from a graph  $G$  the node with the minimum degree was first studied in [2] and shown to compute a 2-approximation of the densest subgraph problem for the  $d_a(G)$  density in [6] and the optimal for the  $d_m(G)$  density in [15].

### 3.2 Solving BFF-AA

To solve the BFF-AA problem, we shall use the average graph  $\hat{H}_G$  of  $\mathcal{G}$ . Lemma 1 shows that  $f_{aa}(S, \mathcal{G})$  is equal with the average density of  $S$  in  $H_G$ . Thus, based on the results of Charikar [6] and Goldberg [8], we conclude that:

**Proposition 2.** The BFF-AA problem can be solved optimally in polynomial time.

Although there exists a polynomial-time optimal algorithms for BFF-AA, the computational complexity of these algorithms (e.g.,  $O(|V||\hat{E}|^2)$  for the case of the max-flow algorithm in [8]), makes them hard to use for large-scale real graphs. Therefore, instead of these algorithm we use the following instantiation of the  $\text{FINDBFF}$ , which we call  $\text{FINDBFF}_A$ . For each node  $v$  in  $S$ , we define its  $\text{score}_a$  in graph history  $\mathcal{G}$  to be equal to its average degree of  $v$ ,

$$\text{score}_a(v, \mathcal{G}[S]) = \frac{1}{|\mathcal{G}|} \sum_{G_t \in \mathcal{G}} \text{degree}(v, G_t[S]).$$

At the  $i$ -th iteration, we select the node  $v_i$  with the *minimum average degree* in  $\mathcal{G}$ . That is,

$$v_i = \arg \min_{v \in S_{i-1}} \text{score}_a(v, \mathcal{G}[S_{i-1}]).$$

Again using Lemma 1 and the results of Charikar [6] we have the following:

**Proposition 3.**  $\text{FINDBFF}_A$  is a  $\frac{1}{2}$ -approximation algorithm for the BFF-AA problem.

*Proof.* It is easy to see that  $\text{FINDBFF}_A$  removes the node with the minimum density in  $\hat{H}_G$ . Charikar [6] has shown that an algorithm that iteratively removes from a graph the node with minimum density provides a  $\frac{1}{2}$ -approximation for finding the subset of nodes that maximizes the average density on a single (weighted) graph snapshot. Given the equivalence we established in Lemma 1,  $\text{FINDBFF}_A$  is also a  $\frac{1}{2}$ -approximation algorithm for BFF-AA.  $\square$

### 3.3 Solving BFF-MA and BFF-AM

We consider the application of  $\text{FINDBFF}_M$  and  $\text{FINDBFF}_A$  algorithms for the two problems. As the following propositions shows, the two algorithms give a poor approximation ratio for both problems. Recall that all our problems are maximization problems, and, therefore, the lower the approximation ratio, the worse the performance of the algorithm.

**Proposition 4.** The approximation ratio of algorithm  $\text{FINDBFF}_M$  for the BFF problem is  $O(\frac{1}{n})$  where  $n$  is the number of nodes.

*Proof.* In the Appendix.  $\square$

**Proposition 5.** The approximation ratio of algorithm  $\text{FINDBFF}_A$  for the BFF-AM problem is  $O(\frac{1}{n})$  where  $n$  is the number of nodes.

*Proof.* In the Appendix.  $\square$

**Proposition 6.** The approximation ratio of algorithm  $\text{FINDBFF}_M$  for the BFF-MA problem is  $O(\frac{1}{\sqrt{n}})$  where  $n$  is the number of nodes.

*Proof.* In the Appendix.  $\square$

We also consider applying the  $\text{FINDBFF}_A$  algorithm that selects to remove the node with the minimum average degree. We can show that  $\text{FINDBFF}_A$  has a poor approximation ratio for the BFF-AM problem.

**Proposition 7.** The approximation ratio of algorithm  $\text{FINDBFF}_A$  for the BFF-MA problem is  $O(\frac{1}{\sqrt{n}})$  where  $n$  is the number of nodes.

*Proof.* In the Appendix.  $\square$

The complexity of BFF-MA and BFF-AM is an open problem and we do not know whether they are solvable in polynomial time or they are NP-hard. Jethava and Beerenwinkel [9] conjecture that the BFF-MA problem is NP-hard, yet they do not provide a proof for this statement.

Given that  $\text{FINDBFF}_A$  and  $\text{FINDBFF}_M$  have no theoretical guarantees, we also investigate a *greedy* approach, which selects which node to remove based on the objective function of the problem at hand. This greedy approach is again an instance of the iterative algorithm shown in Algorithm 1. More specifically, for a target function  $f$  (either  $f_{am}$  or  $f_{ma}$ ), given a set  $S_{i-1}$ , we define the score  $\text{score}_g(v, \mathcal{G}[S_i])$  of node  $v \in S_i$  as follows:

$$\text{score}_g(v, \mathcal{G}[S_{i-1}]) = f(S_{i-1} \setminus \{v\}, \mathcal{G}).$$

At iteration  $i$ , the algorithm selects the node  $v_i$  with the maximum  $\text{score}_g$  to remove, that is,

$$v_i = \arg \max_{v \in S_{i-1}} \text{score}_g(v, \mathcal{G}[S_{i-1}]).$$

We refer to this algorithm as  $\text{FINDBFF}_G$ . Note that the function  $f$  in the scoring function is always the same as the target function  $f$  that is optimized by the algorithm.

**Running time of  $\text{FINDBFF}$ :** The running time of  $\text{FINDBFF}_M$  and  $\text{FINDBFF}_A$  is  $O(n\tau + M)$ , where  $n = |V|$ ,  $\tau$  the number of snapshots in the history graph and  $M = m_1 + m_2 + \dots + m_\tau$  the total number of edges that appear in

all snapshots. To achieve this running time we keep for every snapshot  $G_i$  the list of nodes with degree  $d$ ; these lists can be constructed in time  $O(n\tau)$ . Given this, the time required to find the node with the minimum  $score_m$  (resp.  $score_a$ ) is  $O(\tau)$ . Now in all snapshots, the neighbors of the removed node need to be moved from their position in the  $\tau$  lists; the degree of every neighbor of the removed node is decreased by one. Throughout the execution of the algorithm at most  $O(M)$  such moves can happen. Furthermore, locating the node with the minimum degree in the next iteration of the algorithm can only take time  $O(\tau)$ , as the minimum degree can only decrease by at most one. Therefore, the total running time of the algorithm is  $O(n\tau + M)$ . `FINDBFFG` requires to check all nodes when choosing which node to remove at each step, thus leading to complexity  $O(n^2\tau + nM)$ .

## 4. THE $O^2$ BFF PROBLEM

In this section, we relax the requirement that the nodes are connected in all snapshots of a graph history. Instead, we ask to find the subset of nodes with the maximum aggregate density in at least  $k$  of the snapshots. We call this problem *On-Off BFF* ( $O^2$ BFF) problem.

### 4.1 Problem definition

In the  $O^2$ BFF problem, we seek to find  $k$  graph snapshots from  $\mathcal{G} = \{G_1, \dots, G_\tau\}$  and a set of nodes  $S \subseteq V$  such that the subgraphs induced by  $S$  in all the  $k$  identified snapshots are dense (with the appropriate definition of aggregate density we discussed in Section 2.2). Formally, the  $O^2$ BFF problem is defined as follows:

**Problem 2** (The On-Off BFF ( $O^2$ BFF) Problem). Given a graph history  $\mathcal{G} = \{G_1, G_2, \dots, G_\tau\}$ , an aggregate density function  $f$ , and an integer  $k$ , find a subset of nodes  $S \subseteq V$ , and a subset  $\mathcal{C}_k$  of  $\mathcal{G}$  of size  $k$ , such that  $f(S, \mathcal{C}_k)$  is maximized.

As for Problem 1, depending on the choice of the aggregate density function  $f$ , we have four variants of  $O^2$ BFF. Thus,  $f_{mm}$ ,  $f_{ma}$ ,  $f_{am}$  and  $f_{aa}$  give rise to problems  $O^2$ BFF-MM,  $O^2$ BFF-MA,  $O^2$ BFF-AM and  $O^2$ BFF-AA respectively.

Note that the subcollection of graphs  $\mathcal{C}_k \subset \mathcal{G}$  does not need to consist of contiguous graph snapshots. If this were the case, then the problem could be solved easily by considering all possible contiguous subsets of  $[1, \tau]$  and outputting the one with the highest density. However, all the four variants of the  $O^2$ BFF become NP-hard if we drop the constraint for consecutive graph snapshots.

**Theorem 2.** Problem 2 is NP-hard for any definition of the aggregate density function  $f$ .

*Proof.* For all aggregate density functions, the reduction to the problem is from the  $k$ -CLIQUE problem, but the construction differs slightly depending on the definition of  $f$ . The decision version of the  $k$ -CLIQUE problem given a graph  $G$  asks if the graph contains a clique of size at least  $k$ . The decision version of our problem given a graph history  $\mathcal{G} = \{G_1, \dots, G_\tau\}$  asks if there exists a subset of nodes  $S$  and a subset  $\mathcal{C}_k$  of  $k$  snapshots, such that  $f(S, \mathcal{C}_k) \geq \theta$  for some value  $\theta$ .

We will now describe how to construct the input to our problem for the different density functions. For the case of  $f_{mm}$  and  $f_{am}$  the construction proceeds as follows: given the

graph  $G = (V, E)$ , with  $|V| = n$  nodes, we construct a graph history  $\mathcal{G}$  with  $\tau = n$  snapshots. All snapshots are defined over the vertex set  $V$ . There is a snapshot  $G_i$  for each node  $i \in V$ , consisting of a star-graph with node  $i$  as the center, and edges to all the neighbors of  $i$  in  $G$ .

We will prove that there exists a clique of size at least  $k$  in graph  $G$  if and only if there exists a set of nodes  $S$  and a subset  $\mathcal{C}_k \subseteq \mathcal{G}$  of  $k$  snapshots, with  $f(S, \mathcal{C}_k) \geq 1$ . The forward direction is easy; if there exists a subset of nodes  $S$  in  $G$ , with  $|S| \geq k$ , that form a clique, then selecting this set of nodes  $S$ , and a subset  $\mathcal{C}_k$  of  $k$  snapshots that correspond to nodes in  $S$  will yield  $f_{mm}(S, \mathcal{C}_k) = f_{am}(S, \mathcal{C}_k) = 1$ . This follows from the fact that every snapshot is a complete star where  $d_m(S, G_i) = 1$  for all  $G_i \in \mathcal{C}_k$ . To prove the other direction, we observe that all our snapshots consist of a star graph, and a collection of disconnected nodes. Given a set  $S$ ,  $d_m(S, G_i) = 1$ , if  $i \in S$  and all nodes in  $S$  are connected to the center node  $i$ , and zero otherwise. Therefore, if  $f_{mm}(S, \mathcal{C}_k) = 1$  or  $f_{am}(S, \mathcal{C}_k) = 1$ , then this implies that  $d_m(S, G_i) = 1$  for all  $G_i \in \mathcal{C}_k$ , which means that the  $k$  centers of the graph snapshots in  $\mathcal{C}_k$  are connected to all nodes in  $S$ , and hence to each other. Therefore, they form a clique of size  $k$  in the graph  $G$ .

In the case of  $f_{aa}$  and  $f_{ma}$  the construction proceeds as follows: given the graph  $G = (V, E)$ , with  $|E| = m$  edges, we construct a graph history  $\mathcal{G} = \{G_1, \dots, G_\tau\}$  with  $\tau = m$  snapshots. All snapshots are defined over the vertex set  $V$ . There is a snapshot  $G_e$  for each edge  $e \in E$ , consisting of the single edge  $e$ .

We will prove that there exists a clique of size at least  $k$  in graph  $G$  if and only if there exists a set of nodes  $S$  and a subset  $\mathcal{C}_K \subseteq \mathcal{G}$  of  $K = k(k-1)/2$  snapshots, with  $f(S, \mathcal{C}_K) \geq 1/k$ . The forward direction is easy. If there exists a subset of nodes  $S$  in  $G$ , with  $|S| = k$ , that form a clique, then selecting this set of nodes  $S$ , and the  $\binom{k}{2}$  snapshots  $\mathcal{C}_K$  in  $\mathcal{G}$  that correspond to the edges between the nodes in  $S$  will yield  $f_{aa}(S, \mathcal{C}_K) = f_{ma}(S, \mathcal{C}_K) = 1/k$ .

To prove the other direction, assume that there is no clique of size greater or equal to  $k$  in  $G$ . Let  $\mathcal{C}_K$  be any subset of  $K = k(k-1)/2$  snapshots, and let  $S$  be the union of the endpoints of the edges in  $\mathcal{C}_K$ . Since  $S$  cannot be a clique, it follows that  $|S| = \ell > k$ . Therefore,  $f_{aa}(S, \mathcal{C}_K) = f_{ma}(S, \mathcal{C}_K) = 1/\ell < 1/k$ .  $\square$

### 4.2 Algorithms

We solve all versions of the  $O^2$ BFF problem using a generic algorithm, which we call `FINDO2BFF`, shown in Algorithm 2. `FINDO2BFF` is an iterative algorithm that takes as input a graph history  $\mathcal{G} = \{G_1, \dots, G_\tau\}$ , the aggregate density function  $f$  and an integer  $k$  and outputs a subset of nodes  $S$  and a subset of the snapshots  $\mathcal{C}_k$ . In each iteration, the algorithm starts with some subset of nodes  $S_0 \subseteq V$  and finds the snapshots  $\mathcal{C}_k$  with the top- $k$  highest  $d(G_i[S_0])$  score; this is done by the `BESTSNAPSHOTS` routine. `BESTSNAPSHOTS` computes the density  $d(G_i[S_0])$  of  $S_0$  in all snapshots  $G_i \in \mathcal{G}$  and outputs the  $k$  snapshots  $\mathcal{C}_k$  with the highest density. Given  $\mathcal{C}_k$  the algorithm then finds the set  $S \subseteq V$  such that  $f(\mathcal{C}_k, S)$  is maximized. This step essentially solves Problem 1 on input  $\mathcal{C}_k$  and aggregate density function  $f$ ; thus this step is solved using the `FINDBFF` algorithm. The `FINDO2BFF` algorithm stops when it finds a set of snapshots  $\mathcal{C}_k$  and a set of nodes  $S$  such that no further iterations can improve the score  $f(\mathcal{C}_k, S)$ .

Depending on whether we are solving the  $O^2\text{BFF-MM}$ ,  $O^2\text{BFF-MA}$ ,  $O^2\text{BFF-AM}$  or  $O^2\text{BFF-AA}$  problems, we use the appropriate version of the  $\text{FINDBFF}$  algorithm. The version of  $\text{FINDO}^2\text{BFF}$  is specified by the aggregate density function  $f$ , which is used as an argument both in the  $\text{FINDBFF}$  routine, and by the version of the  $\text{FINDBFF}$  algorithm we use (i.e.,  $\text{FINDBFF}_M$ ,  $\text{FINDBFF}_A$ , or  $\text{FINDBFF}_G$ ). Thus we get three different versions of  $\text{FINDO}^2\text{BFF}$ :  $\text{FINDO}^2\text{BFF}_M$ ,  $\text{FINDO}^2\text{BFF}_A$  and  $\text{FINDO}^2\text{BFF}_G$  respectively.

---

**Algorithm 2** The  $\text{FINDO}^2\text{BFF}$  algorithm.

---

**Input:** Graph history  $\mathcal{G} = \{G_1, \dots, G_\tau\}$ ; an aggregate-density function  $f$ ; integer  $k$   
**Output:** A subset of nodes  $S$  and a subset of snapshots  $\mathcal{C}_k \subseteq \mathcal{G}$ .

---

```

1: converged = False
2:  $(\mathcal{C}^0, S^0) = \text{INITIALIZE}(\mathcal{G}; f)$ 
3:  $F^0 = 0$ 
4: while not converged do
5:    $\mathcal{C}_k = \text{BESTSNAPSHOTS}(S^0, f)$ 
6:    $S = \text{FINDBFF}(\mathcal{C}_k; f)$ 
7:    $F = f(\mathcal{C}_k, S)$ 
8:   if  $F < F^0$  then
9:     Converged = True
10:  else  $F^0 = F$ ,  $S^0 = S$ 
11:  end if
12: end while
13: return  $S, \mathcal{C}_k$ 

```

---

An important step of the  $\text{FINDO}^2\text{BFF}$  is the initialization step (routine  $\text{INITIALIZE}$ ). We experiment with three different alternatives for this initialization: *random*, *contiguous* and *at least- $k$* . We discuss these alternatives below:

**Random initialization:** For this initialization we randomly pick  $k$  snapshots  $\mathcal{C}^0$  from  $\mathcal{G}$ . These snapshots are then used for solving the corresponding BFF problem on input  $\mathcal{C}^0$  and produce  $S^0 = \text{FINDBFF}(\mathcal{C}_k; f)$ .

**Contiguous initialization:** This initialization first finds an  $S^0$  that is the best subset for a contiguous set of snapshots. Thus, given  $\mathcal{G} = \{G_1, \dots, G_\tau\}$ , this initialization technique goes over all the  $O(\tau)$  contiguous sets of  $k$  snapshots from  $\mathcal{G}$ , and finds the set of  $k$  snapshots  $\mathcal{C}^0$  and set of nodes  $S^0$  such that:  $f(S^0, \mathcal{C}^0)$  is maximized. This set  $S^0$  is then used for the rest of the steps of the algorithm. The intuition behind this initialization technique is that it assumes that the best  $k$  snapshots of  $\mathcal{G}$  are going to be contiguous. Our experiments demonstrate that in practice this is true in many datasets – e.g., in collaboration networks that evolve over time and we expect to see some temporal locality.

**At least- $k$  initialization:** For this initialization, we solve the BFF problem independently in each snapshot  $G_i \in \mathcal{G}$ . This results into  $\tau$  different sets  $S_i \subseteq V$ , one for each solution of BFF on  $G_i$ .  $S^0$  includes the nodes that appear in at least  $k$  of the  $\tau$  sets  $S_i$ . The intuition behind this initialization is to include in the initial solution those nodes that appear to be densely connected in many snapshots. We also experimented with other natural alternatives, such as the union:  $S^0 = \bigcup_{i=1 \dots \tau} S_i$  and the intersection:  $S^0 = \bigcap_{i=1 \dots \tau} S_i$ ; the at least- $k$  approach seems to strike a balance between the two.

**Running time:** Observe that independently of the initialization method the  $\text{INITIALIZE}$  routine takes time less than the running time of  $\text{FINDBFF}$ , i.e.,  $O(n\tau + M)$ . Similarly the running time of the  $\text{BESTSNAPSHOTS}$  routine is also less than  $O(n\tau + M)$ . Therefore, the running time of the  $\text{FINDBFF}$  algorithm is  $O(I(n\tau + M))$ , where  $I$  is the number of iterations required until convergence. In practice, we observed that independently of the version of the  $O^2\text{BFF}$  problem that  $\text{FINDO}^2\text{BFF}$  is trying to solve, it converges in at most 6 iterations.

## 5. BFF PROBLEM EXTENSIONS

The definitions of the BFF and  $O^2\text{BFF}$  problem focus on the identification of a set of nodes  $S$  such that their aggregate density is maximized. We now consider natural extensions of the BFF problem by placing additional constraints on the dense subgraphs.

**Query-node constraint:** An interesting extension is introducing a set  $Q$  of *seed query nodes* and requiring that the output set of nodes  $S$  has high density and also contains the input seed nodes. A similar extension was introduced for static (e.g., single snapshots) graphs in [15]. In practice, this variant of BFF identifies the lasting “best friends” of the query nodes. We call this the QR-BFF problem.

We can modify the  $\text{FINDBFF}$  algorithms appropriately so that they take into consideration this additional constraint. In particular,  $\text{FINDBFF}_M$  stops when a query node in  $Q$  is selected to be removed. Let us call this modified algorithm,  $\text{QR-FINDBFF}_M$ . We can prove the following proposition. (We omit the proof due to space constraints.)

**Proposition 8.**  $\text{QR-FINDBFF}_M$  solves the QR-BFF-MM problem optimally in polynomial time.

We also modify  $\text{FINDBFF}_A$  so that it does not remove seed nodes as follows: If at any step, the node with the minimum average degree happens to be a seed node, the algorithm selects to remove the node with the next smallest degree that is not a seed node. The algorithm stops when the only remaining nodes are seed nodes. Let us call this modified algorithm,  $\text{QR-FINDBFF}_A$ .

**Proposition 9.** Let  $S^*$  be an optimal solution for the QR-BFF-AA problem and  $S_A$  be the solution of the  $\text{QR-FINDBFF}_A$  algorithm. It holds:  $f_{aa}(S_A) \geq \frac{s f_{aa}(S^*) + 2\omega}{2(s+q)}$ , where  $q = |Q|$ ,  $s = |S^* \setminus Q|$  and  $\omega = \sum_{u \in Q} \text{degree}(u, S^*)$ .

*Proof.* By Lemma 1, it suffices to show that the  $\text{QR-FINDBFF}_A$  algorithm provides an approximation of the average density of a single graph  $G$ . Let  $S^*$  be the optimal solution for  $G$ . Let  $G'$  be the graph that results from  $G$  when we delete all edges between two query nodes in  $G$ . Clearly,  $S^*$  is also an optimal solution for  $G'$ . Assume that we assign each edge  $(u, v)$  to either  $u$  or  $v$ . For each node  $u$ , let  $a(u)$  be the number of edges assigned to it and let  $a_{max} = \max_u \{a(u)\}$ . It is easy to see that  $f_{aa}(S^*) \leq \frac{1}{2} a_{max}$ , since each edge in the optimal solution must be assigned to a node in it. Now assume that the assignment of edges to nodes is performed as the  $\text{QR-FINDBFF}_A$  algorithm proceeds. Initially, all edges are unassigned. When at step  $i$ , a node  $u$  is deleted, we assign to  $u$  all the edges that go from  $S_{i-1}$  to  $u$ . Note that this assignment maintains the invariant that at each step, all edges between two nodes in the current set  $S$  are unassigned, while all other edges are assigned. When the

algorithm stops, all edges have been assigned. Consider a single iteration of the algorithm when a node  $u_{min}$  is selected to be removed and let  $S$  be the current set. Let  $s$  be the number of non-query nodes in  $S$ , and  $q = |Q|$  be the number of query nodes. It holds:  $f_{aa}(S) = \frac{1}{s+q} \sum_{u \in S} \text{degree}(u) = \frac{1}{s+q} \sum_{v \in S \setminus Q} \text{degree}(u) + \frac{1}{s+q} \sum_{u \in Q} \text{degree}(u)$ . Let  $\Omega = \frac{1}{s+q} \sum_{u \in Q} \text{degree}(u)$ . Since  $u_{min}$  has the smallest degree among all nodes in  $S$  but the seed nodes, we have  $f_{aa}(S) \geq \frac{1}{s+q} s a(u_{min}) + \Omega$ . Since all edges are assigned and edges are assigned to a node only when this node is removed, at some step of the execution of the algorithm  $a(u_{min}) = a_{max}$ . Thus, for some  $S$ ,  $f_{aa}(S) \geq \frac{s}{s+q} a_{max} + \Omega \geq \frac{s}{s+q} \frac{1}{2} f_{aa}(S^*) + \Omega$ .  $\square$

**Connectivity constraint:** Another meaningful extension is to impose restrictions on the connectivity of  $S$ . The connectivity of  $S$  in a graph history  $\mathcal{G} = \{G_1, \dots, G_\tau\}$  may have many different interpretations. One may consider a version where all the induced subgraphs  $G_t[S]$  for  $t \in \{1, \dots, \tau\}$  are connected. Another alternative is that at least  $m > 0$  of the  $\tau$   $G_t[S]$ 's are connected. Here, we assume that a definition of connectivity for  $S$  is given in the form of a predicate  $\text{connected}(S, \mathcal{G})$  which is true if  $S$  is connected and false otherwise. Our problem now becomes: given a graph history  $\mathcal{G}$ , a set of query nodes  $Q \subset V$  and an aggregate density function  $f$ , find a subset of nodes  $S \subseteq V$ , such that (1)  $f(S, \mathcal{G})$  is maximized, (2)  $Q \subseteq S$ , and (3)  $\text{connected}(S, \mathcal{G})$  is true.

To solve this problem, we can modify `FINDBFF` so that it tests for the connectivity predicate  $\text{connected}(S, \mathcal{G})$  and stops when the connectivity constraint no longer holds. In our experiments, we apply a simplest test, just running the algorithms on the connected components of the query nodes.

**Size constraint:** Finally, note that the definition of BFF does not impose a constraint on the size of the output set of nodes  $S$ . In that respect the problem is parameter-free. If necessary, one can add an additional constraint to the problem definition by imposing a cardinality constraint on the output  $S$ . However the cardinality constraint makes the subgraph-discovery problem computationally hard [15]. This also holds for the BFF problem; simply consider a graph history with replicas of the same single-snapshot graph.

## 6. EXPERIMENTAL EVALUATION

In this section, we provide an evaluation of our approach. The goal of our experimental evaluation is threefold. First, we want to evaluate the performance of our algorithms for the BFF and the  $O^2$ BFF problems in terms of the quality of the solutions and running time. Second, we want to compare the solutions for the different variants of the aggregate density functions. Third, we want to show the usefulness of the problem, by presenting results of BFFs and  $O^2$ BFFs with and without seed query nodes in two real datasets, namely research collaborators in *DBLP* and hashtags in *Twitter*.

We ran our experiments on a system with a quad-core Intel Core i7-3820 3.6 GHz processor, with 64 GB memory. We only used one core in all experiments.

**Datasets.** To evaluate our algorithms, we use a number of real graph histories, where the snapshots correspond to collaboration, social and computer networks.

Table 1: Real dataset characteristics

Dataset	# Nodes	# Edges (aver. per snapshot)	# Snapshots
<i>DBLP</i> <sub>10</sub>	2,625	1,143	10
<i>DBLP</i> <sub>5</sub>	1,327	872	5
<i>Oregon</i> <sub>2</sub>	11,806	31,559	9
<i>Slashdot</i>	82,144	488,902	3
<i>Caida</i>	31,379	45,833	122
<i>AS</i>	7,716	7,783	733
<i>Twitter</i>	91,150	100	15

- In the *DBLP*<sup>1</sup> datasets, each snapshot corresponds to a year. The nodes are authors and there is an edge between two authors in a graph snapshot, if they co-authored a paper in the corresponding year. The dataset includes papers published in 11 top database and data mining conferences. We use two datasets: *DBLP*<sub>10</sub> that contains publications in the [2006, 2015] interval and *DBLP*<sub>5</sub> that contains publications in the [2011, 2015] interval. We only consider as co-authors, authors that wrote at least two papers together in the corresponding interval.
- The *Caida*<sup>2</sup> dataset, contains 122 CAIDA autonomous systems (AS) graphs, derived from a set of route views BGP-table instances.
- The *AS*<sup>3</sup> dataset represents a communication network of who-talks-to-whom from the BGP (Border Gateway Protocol) logs. The dataset contains 733 daily snapshots which span an interval of 785 days from November 8, 1997 to January 2, 2000.
- The *Oregon*<sub>2</sub><sup>4</sup> dataset consists of nine AS graphs, one snapshot per week between March 31, 2001 and May 26, 2001. Each snapshot represents AS peering information inferred from Oregon route-views, looking glass data, and Routing registry, all combined.
- The *Slashdot*<sup>5</sup> dataset consists of nodes who are users of the Slashdot Zoo social network and edges who are friend/foe links between the users in different snapshots.
- In the *Twitter* dataset [16], nodes are hashtags of tweets and edges represent the co-appearance of hashtags in a tweet. The dataset contains 15 daily snapshots from October 27, 2013 to November 10, 2013.

The dataset characteristics are summarized in Table 1.

### 6.1 Evaluation of the BFF algorithms and aggregate density variants

In the first set of our experiments, we run the `FINDBFF` algorithms for the BFF-MM, BFF-MA, BFF-AM and BFF-AA problems using all our real datasets and present the results in Table 2, where we report the value of the objective function achieved by each algorithm and the size of the solution.

Since, as shown in Section 3, `FINDBFFM` and `FINDBFFA` are provably good for the BFF-MM and BFF-AA problems respectively, these are the only algorithms used for these problems. For solving the BFF-MA and BFF-AM problems, we use all three algorithms; i.e., `FINDBFFM`, `FINDBFFA` as well as `FINDBFFG`. For the BFF-MA problem, we also use

<sup>1</sup><http://dblp.uni-trier.de/>

<sup>2</sup><http://www.caida.org/data/as-relationships/>

<sup>3</sup><https://snap.stanford.edu/data/as.html>

<sup>4</sup><https://snap.stanford.edu/data/oregon2.html>

<sup>5</sup><https://snap.stanford.edu/data/soc-sign-Slashdot090221.html>

Table 2: Results of the the different algorithms for the BFF problem on the real datasets.

Datasets	BFF-MM		BFF-MA								BFF-AM								BFF-AA	
	FINDBFF <sub>M</sub>		FINDBFF <sub>M</sub>		FINDBFF <sub>A</sub>		FINDBFF <sub>G</sub>		DCS		FINDBFF <sub>M</sub>		FINDBFF <sub>A</sub>		FINDBFF <sub>G</sub>		FINDBFF <sub>A</sub>		FINDBFF <sub>A</sub>	
	Size	$f_{mm}$	Size	$f_{ma}$	Size	$f_{ma}$	Size	$f_{ma}$	Size	$f_{ma}$	Size	$f_{am}$	Size	$f_{am}$	Size	$f_{am}$	Size	$f_{am}$	Size	$f_{aa}$
<i>DBLP<sub>10</sub></i>	11	1.0	3	1.33	8	1.75	61	1.7	14	1.29	11	1.0	4	1.7	4	1.0	8	2.75		
<i>DBLP<sub>5</sub></i>	26	2.0	26	2.15	4	2.5	30	2.27	4	2.5	26	2.0	4	2.8	6	1.0	12	3.07		
<i>Oregon<sub>2</sub></i>	75	23.0	140	44.33	131	45.24	132	45.95	116	44.91	63	24.44	44	23.22	461	3.22	147	47.89		
<i>Slashdot</i>	24	5.0	6,246	23.16	4,749	43.3	4,617	43.33	5,415	23.57	14	5.33	92	10.67	115	14.33	4,181	44.81		
<i>Caida</i>	17	8.0	33	13.76	29	12.76	60	15.43	57	15.05	20	12.72	36	18.11	311	3.43	96	33.21		
<i>AS</i>	15	4.0	19	8.53	18	6.67	20	9.0	16	8.75	12	7.44	14	9.05	91	3.04	38	16.38		
<i>Twitter</i>	-	0.0	836	0.04	7	0.29	13	0.62	720	0.05	-	0.0	3	1.0	3	1.0	5	1.38		

Table 3: Relative overlap of the solutions of the different problems for the real datasets.

<i>DBLP<sub>10</sub></i>	BFF-MM	BFF-MA	BFF-AM	BFF-AA
BFF-MM	1.00	1.00	0.27	0.36
BFF-MA	1.00	1.00	0.27	0.36
BFF-AM	0.75	0.75	1.00	1.00
BFF-AA	0.80	0.80	0.80	1.00
<i>Oregon<sub>2</sub></i>	BFF-MM	BFF-MA	BFF-AM	BFF-AA
BFF-MM	1.00	1.00	0.56	1.00
BFF-MA	0.54	1.00	0.31	0.87
BFF-AM	0.95	1.00	1.00	1.00
BFF-AA	0.51	0.83	0.30	1.00
<i>Caida</i>	BFF-MM	BFF-MA	BFF-AM	BFF-AA
BFF-MM	1.00	1.00	1.00	1.00
BFF-MA	0.52	1.00	0.82	1.00
BFF-AM	0.47	0.75	1.00	1.00
BFF-AA	0.18	0.34	0.38	1.00

the *DCS* algorithm proposed in [9] for a problem similar to BFF-MA. The *DCS* algorithm is also an iterative algorithm that removes one node at a time. At each step, *DCS* finds the subgraphs with the largest average density for each of the snapshots. Then, it identifies the subgraph with the smallest average density among them and removes the node that has the smallest degree in this subgraph.

**Comparison of the algorithms for BFF-MA and BFF-AM :** As shown in Table 2, for the BFF-MA problem, FINDBFF<sub>M</sub> and *DCS* have comparable performance, since they both remove nodes with small degrees. FINDBFF<sub>A</sub> outperforms both of them, since in all but the *Caida* and *AS* datasets, FINDBFF<sub>A</sub> returns subgraphs with larger aggregate density values. In the *Caida* and *AS* datasets, due probably to their large number of snapshots, FINDBFF<sub>A</sub> – which is based on the average degree – returns dense subgraphs with slightly smaller density than FINDBFF<sub>M</sub> and *DCS*. FINDBFF<sub>G</sub> performs the best in all datasets.

When comparing the performance of FINDBFF<sub>M</sub>, FINDBFF<sub>A</sub> and FINDBFF<sub>G</sub> for BFF-AM, we can clearly see that FINDBFF<sub>A</sub> achieves the best value of the objective function. Our deeper analysis of the inferior performance of FINDBFF<sub>G</sub> for this problem revealed that FINDBFF<sub>G</sub> often gets trapped in local maxima after removing just a few nodes of the graph and it cannot find good solutions.

Regarding execution time, all algorithms required from a few milliseconds to a few seconds to produce a solution in all datasets.

**Comparison of aggregate densities:** Looking at the results of Table 2, we observe that the solutions for BFF-MM usually have small cardinality compared to the solutions for

other problems. This is due to the fact that there are not so many groups of nodes with large  $f_{mm}$  value, since this objective is rather strict. As a result, the solutions reported for this problem are usually small (groups of) cliques. The solutions for the same problem in the autonomous-system datasets appear to have higher  $f_{mm}$  scores. This may be due to the fact that there are larger groups of nodes with lasting connections in these datasets, e.g., nodes that communicate intensely between each other during the observation period.

Note that the good solutions to BFF-MA (e.g., those obtained by FINDBFF<sub>A</sub>) appear to have the largest cardinality among the solutions for all other problems. This is because in order to maximize  $f_{ma}$ , the algorithm needs to find a solution that maximizes the minimum average degree across snapshots. The more nodes a solution includes the more likely it is to maximize this value. Finally, as expected, the value of the aggregate density of the reported solution (independently of the problem variant) increases with the density of the graphs.

**Overlap.** We perform an additional experiment to explore the similarity between the actual sets of nodes reported by the FINDBFF algorithm for each one of the four BFF problem variants. The results of these comparisons are shown in Table 3. The entries of each subtable that corresponds to a single dataset should be interpreted as follows: the row  $r$  column  $c$  entry of the subtable is a value in  $[0, 1]$ . This value is computed by taking the intersection between the solutions obtained for the problems depicted in  $r$  and  $c$  divided by the cardinality of the solution of the problem in row  $r$ . In all cases, the solutions to the problems are obtained using the appropriate FINDBFF algorithm.

The results of Table 3 show that there are cases where there is overlap between the solutions reported for the different problems. Moreover, we observe that usually the solutions to BFF-MM are subsets of the solutions to BFF-MA. Furthermore, the solutions to BFF-AA is a superset of the solutions to all other problems across datasets.

**Using synthetic data:** Finally, we evaluate the accuracy of our algorithms in discovering planted dense subgraphs in graph histories. Since we do not have any ground-truth information for real data, we conduct this experiment using synthetic datasets. We generate the data for this experiment as follows. First, we create 10 graph snapshots with 4000 nodes using the Forest Fire model [11] with the default forward and backward burning probabilities of 0.35. Then, in each one of the 10 snapshots we plant a dense random subgraph  $A$  that has 100 nodes and edge probability  $p_A = 0.5$ . Note, that  $A$  has different edges at different snapshots. We also add a second dense subgraph  $B$  with the same number of nodes as  $A$  and edge probability  $p_B = \{0.5, 0.7, 0.9\}$ . We plant  $B$  in  $\ell$  random snapshots, for different values of  $\ell$ . Us-



Table 4: The F-measure for synthetic datasets.

# $\ell$	$P_B = 0.5$				$P_B = 0.7$				$P_B = 0.9$			
	BFF-MM	BFF-MA	BFF-AM	BFF-AA	BFF-MM	BFF-MA	BFF-AM	BFF-AA	BFF-MM	BFF-MA	BFF-AM	BFF-AA
1	1.0	1.0	1.0	1.0	0.99	1.0	1.0	1.0	0.97	1.0	1.0	1.0
3	0.98	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.98	1.0	1.0	1.0
5	0.99	1.0	1.0	1.0	0.99	1.0	0.99	1.0	0.97	1.0	1.0	1.0
7	0.99	1.0	1.0	1.0	1.0	1.0	0.99	0.66	1.0	1.0	0.0	0.0
9	0.98	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.99	1.0	0.0	0.0

ing the resulting history graph, we run our algorithm to find the solution that maximizes the aggregate density for each input. Then, check whether the solutions to the BFF-MM, BFF-MA, BFF-AM and BFF-AA problems (as found by the appropriate version of the FINDBFF algorithm) can identify  $A$  as the dense set of nodes they report.

In Table 4, we report the  $F$  measure of the obtained solutions; recall that the  $F$  measure takes values in  $[0, 1]$  and the larger the value the better the recall and precision of the solution with respect to the ground truth (in this case  $A$ ). Our experiments indicate that our algorithms find subgraph  $A$  in almost all cases. Our algorithms achieve  $F$  values close to 1 in all cases, even when  $\ell = 9$  and  $p_B = 0.9$ , except from the problems BFF-AM and BFF-AA where the value of the  $F$  measure drops as  $p_B$  and  $\ell$  increase. This once again reveals the “strictness” of the  $f_{mm}$  and  $f_{ma}$  functions in evaluating dense subgraphs in graph histories.

## 6.2 Evaluation of the O<sup>2</sup>BFF algorithms

In this set of experiments, we evaluate the performance of the FINDO<sup>2</sup>BFF algorithms. We applied the FINDO<sup>2</sup>BFF algorithm on all real datasets for various values of  $k$ . In Figures 2 and 3, we report results (i.e., the value of the aggregate density) for two of them, namely *DBLP* and *Oregon2* for different values of  $k$ , which is expressed as a percentage of the total number of snapshots of the input graph history.

Overall, we observe that the *contiguous* initialization is slightly better than the rest in many cases. This is indicative of *temporal locality* of dense subgraphs in some datasets, i.e., in these datasets dense subgraphs are usually alive in a few contiguous snapshots. This is very evident in datasets from collaboration networks such as the *DBLP* datasets. When comparing the results across all problems, we notice that the solutions of FINDO<sup>2</sup>BFF<sub>A</sub> for O<sup>2</sup>BFF-AA is the *least* sensitive to the initialization method used. We also observe that as  $k$  increases the aggregate density of the solutions decrease. This again is explained by the fact that often dense subgraphs are only “alive” in a few snapshots.

*Convergence and running time:* In terms of convergence, FINDO<sup>2</sup>BFF requires 2-6 iterations to converge in all datasets. The total execution time for producing a solution for the small datasets was a few milliseconds, whereas for dense graphs, such as *Slashdot*, the algorithm required 2-3 minutes – depending on the value of  $k$ .

**Using synthetic data:** We also evaluate the performance of FINDO<sup>2</sup>BFF using synthetic datasets. In this experiment, we use the same graph generation setting as discussed before, with the only difference that we now plan the dense random subgraph  $A$  in  $\ell \in \{2, 4, 6, 8\}$  of the 10 snapshots and run the FINDO<sup>2</sup>BFF algorithm with  $k = \ell$ . We plant the dense subgraph either in contiguous snapshots, or in random ones and report the achieved aggregate density in Figures 4 and 5 respectively. With contiguous planting, the contigu-

Table 5: The authors output as solutions to the BFF problem on *DBLP*<sub>10</sub> (in parentheses the dense subgraphs).

BFF-MM
(Wei Fan, Philip S. Yu, Jiawei Han, Charu C. Aggarwal), (Lu Qin, Jeffrey Xu Yu, Xuemin Lin), (Guoliang Li, Jianhua Feng), (Craig Macdonald, Iadh Ounis)
BFF-MA
(Wei Fan, Jing Gao, Philip S. Yu, Jiawei Han, Charu C. Aggarwal), (Jeffrey Xu Yu, Xuemin Lin, Ying Zhang)
BFF-AM
(Wei Fan, Jing Gao, Philip S. Yu, Jiawei Han)
BFF-AA
(Wei Fan, Jing Gao, Philip S. Yu, Jiawei Han, Charu C. Aggarwal, Mohammad M. Masud, Latifur Khan, Bhavani M. Thuraisingham)

ous initialization finds the dense subgraph for all  $k$ , while the at-least- $k$  initialization works well for large  $k$ . For random planting, the at-least- $k$  initialization works well, with the exception of O<sup>2</sup>BFF-MM for small  $k$  where there are outliers, i.e., other nodes outside  $A$  with large degrees. The contiguous initialization improves as  $k$  increases, since in this case, some of the snapshots become inevitably contiguous.

## 6.3 Case studies

In this section, we report indicative results we obtained using the *DBLP*<sub>10</sub> and the *Twitter* datasets. These results identify lasting dense author collaborations and hashtag co-occurrences respectively.

**Lasting dense co-authorships in *DBLP*<sub>10</sub>:** In Table 5 we report the set of nodes output as solutions to BFF-MM, BFF-MA, BFF-AM and BFF-AA on the *DBLP*<sub>10</sub> dataset. First observe that BFF-MM includes many small subgraphs, while BFF-AA includes one large subgraph. Moreover, three authors “Wei Fan”, “Philip S. Yu”, and “Jiawei Han” are part of *all* four solutions. Despite the fact that these authors have co-authored only two papers together in our dataset, subsets of them have collaborated very frequently over the last decade. Interestingly, the solution for BFF-AA contains a superset of collaborators of these authors. Although this set of authors never appears to have been co-authors in the same paper simultaneously, subset of the authors have collaborated with each other in many snapshots. Thus, including the extra authors increased the value of  $f_{aa}$ . Although the group “Wei Fan”, “Philip S. Yu”, “Jiawei Han” and “Charu C. Aggarwal” also participates in the solutions of BFF-MM and BFF-MA, in these solutions we also see new names. These are authors that have no collaborations with the former group, but they form a dense subgraph within themselves. Thus, the solution reported for BFF-MM and BFF-MA consists of more than one dense subgraphs.

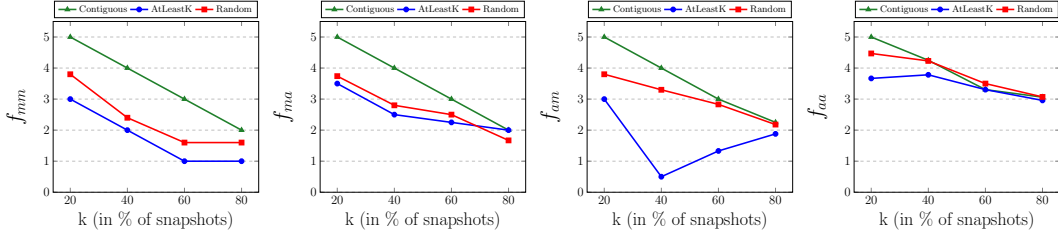


Figure 2: DBLP<sub>10</sub> dataset: scores of aggregate density functions  $f$

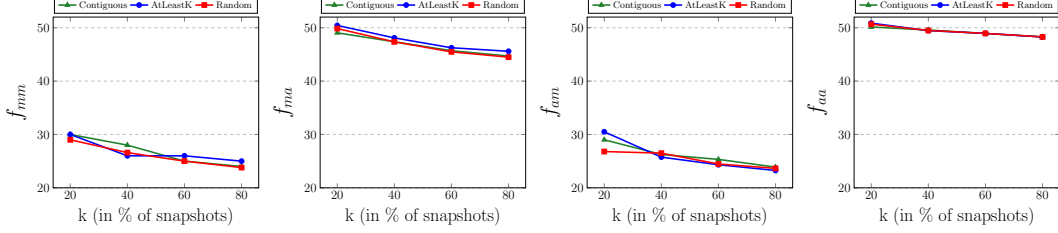


Figure 3: Oregon<sub>2</sub> dataset: scores of aggregate density functions  $f$

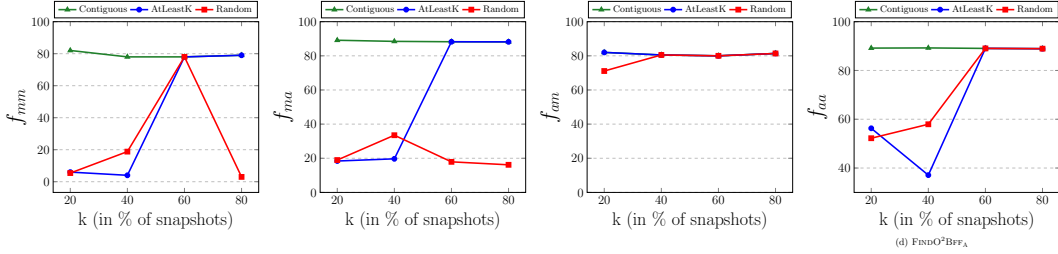


Figure 4: Synthetic dataset with contiguous planting: scores of aggregate density functions  $f$

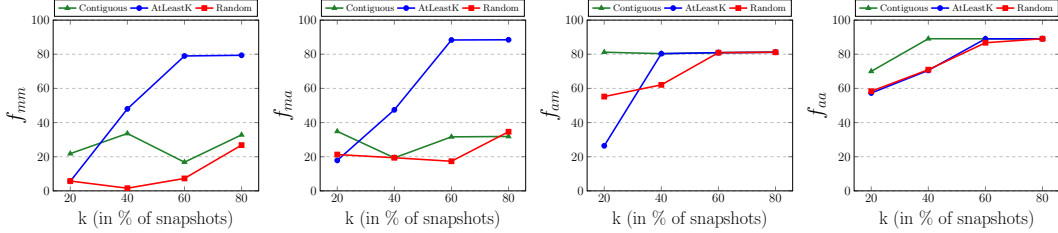


Figure 5: Synthetic dataset with random planting: scores of aggregate density functions  $f$

In Table 6, we report results for O<sup>2</sup>BFF-MM, O<sup>2</sup>BFF-MA, O<sup>2</sup>BFF-AM and O<sup>2</sup>BFF-AA on the same dataset. These authors are the most dense collaborators for  $k = 2, 4, 6$ , and 8 (recall there are 10 years in the dataset). We also report the corresponding years of their dense collaborations. Many new groups of authors appear. For example, we have new groups of collaborators from Tsinghua University, CMU and RPI among others. The authors appeared in the solutions of BFF also appear here for large values of  $k$ .

We also studied experimentally the QR-BFF problem. In Table 7, we show indicative results for three of the authors of this paper as seed nodes. For *E. Pitoura*, we retrieve a group of ex-graduate students with whom she had a lasting and prolific collaboration; for *E. Terzi* close collaborators from BU University, and for *P. Tsaparas*, a group of collaborators from his time at Microsoft Research. Note that in the last case, the selected set consists of researchers with

whom *P. Tsaparas* has co-authored several papers in the period recorded in our dataset, but these authors are also collaborating amongst themselves. Finally, we use one of the authors appearing in the dense subgraphs of the O<sup>2</sup>BFF, namely *C. Faloutsos* as seed node. In this case, we obtain a dense subgraph similar to the one we have reported in Table 6. Finally, we consider a query with two authors: *C. Faloutsos* and his student *D. Koutra*. Adding *D. Koutra* to the query set changes the consistency of the result, focusing more on authors that are collaborators of both query nodes.

**Lasting dense hashtag appearances in Twitter:** In Table 8, we report results of the O<sup>2</sup>BFF problem on the *Twitter* dataset. Note that the results of the BFF problem on this dataset (as shown in Table 2) are very small graphs, since very few hashtags appear together in all 15 days of the dataset. As seen in Table 8, we were able to discover

Table 6: The authors output as solutions to the  $O^2BFF$  problem on  $DBLP_{10}$ .

<b>k = 2</b>	BFF-MM, BFF-MA, BFF-AM, BFF-AA (Christos Faloutsos, Leman Akoglu, Lei Li, Keith Henderson, Hanghang Tong, Tina Eliassi-Rad) Years: 2010 - 2011			
<b>k = 4</b>	BFF-MM, BFF-MA, BFF-AM (Mo Liu, Chetan Gupta, Song Wang, Ismail Ari, Elke A. Rundensteiner) Years: 2010 - 2013		BFF-AA (Yong Yu, Dingyi Han, Zhong Su, Lichun Yang, Shengliang Xu, Shenghua Bao) 2007, 2009 - 2011	
<b>k = 6</b>	BFF-MM, BFF-MA, BFF-AM (Liyun Ru, Min Zhang, Yiqun Liu, Shaoping Ma) Years: 2007 - 2012		BFF-AA (Min Zhang, Liyun Ru Bhavani, Yiqun Liu, Shaoping Ma), (Latifur Khan, M. Thuraisingham, Mohammad M. Masud, Wei Fan, Jing Gao, Philip S. Yu, Jiawei Han) 2007 - 2012	
<b>k = 8</b>	BFF-MM (Min Zhang, Yiqun Liu, Shaoping Ma) Years: 2007 - 2014	BFF-MA (Wei Fan, Jing Gao, Philip S. Yu, Jiawei Han, Charu C. Aggarwal) 2007 - 2008, 2010 - 2015	BFF-AM (Liyun Ru, Min Zhang, Yiqun Liu, Shaoping Ma) 2007 - 2014	BFF-AA (Latifur Khan, Bhavani M. Thuraisingham, Mohammad M. Masud, Wei Fan, Jing Gao, Philip S. Yu, Jiawei Han, Charu C. Aggarwal) 2007 - 2012, 2014 - 2015

Table 7: An example of authors output as solutions to the QR-BFF problem on  $DBLP_{10}$ 

QR-BFF
<b>E. Pitoura:</b> G. Koloniari, M. Drosou, K. Stefanidis
<b>E. Terzi:</b> V. Ishakian, D. Erdos, A. Bestavros
<b>P. Tsaparas:</b> A. Fuxman, A. Kannan, R. Agrawal
<b>C. Faloutsos, D. Koutra:</b> Chris H. Q. Ding, L. Akoglu, H. Huang, Lei Li, Tao Li, H. Tong

interesting dense subgraphs of hashtags appearing in  $k = 3$ , 6, and 9 of these days. These hashtags are related to various events (including fl races and wikileaks). Note also, that for large values of  $k$ , we do not get interesting results which is a fact consistent with the ephemeral nature of Twitter, where hashtags are short-lived. This is especially true for  $f_{mm}$  and  $f_{ma}$  that are more strict in that they impose density constraints in each and every snapshot.

## 7. RELATED WORK

To the best of our knowledge, we are the first to systematically study all the variants of the BFF, and  $O^2BFF$  problems.

The research most related to ours is the recent work of Jethava and Beerenwinkel [9] and Rozenshtein *et al.* [14]. To the best of our understanding, the authors of [9] introduce one of the four variants of the BFF problem we studied here, namely, BFF-MA. In their paper, the authors conjecture that the problem is NP-hard and they propose a heuristic algorithm. Our work performs a rigorous and systematic study of the general BFF problem for multiple variants of the aggregate density function. Additionally, we introduce and study the  $O^2BFF$  and QR-BFF problems, which were not studied in [9]. The recent work in [14] focuses on the discovery of dense subgraphs over graph snapshots. This problem is more similar to the  $O^2BFF$  version of our problem as their goal is to identify an interval and a subset of nodes such that these nodes are dense in the graph consisting of the *union* of edges appearing in any snapshot of the interval.

Furthermore, the authors focus on the identification of a *contiguous* interval. In our case, the  $k$  snapshots reported as part of the solution to  $O^2BFF$  are not necessarily contiguous.

We review next work on discovering dense subgraphs in single static graphs and evolving graphs.

**Dense subgraphs of static graphs:** There is a huge literature on extracting “dense” subgraphs from a single graph snapshot. Most formulations for finding subgraphs that define near-cliques are often NP-hard and often hard to approximate due to their connection to the maximum-clique problem [1, 5, 12, 13, 17]. As a result, the problem of finding the subgraph with the maximum average or minimum degree has become particularly popular, due to its computational tractability. Specifically, the problem of finding a subgraph with the maximum average degree can be solved optimally in polynomial time [6, 8, 10], and there exists a practical greedy algorithm that gives a 2-approximation guarantee in time linear to the number of edges and nodes of the input graph [6]. The problem of identifying a subgraph with the maximum minimum degree, can be solved optimally in polynomial time [15], using again the greedy algorithm proposed by Charikar [6]. In our work, we use the average and minimum degree as ways to quantify the density of the subgraph in a single graph snapshot, and we extend these definitions to sets of snapshots. The algorithmic techniques we use for the BFF problem are inspired by the techniques proposed by Charikar [6], and by Sozio and Gionis [15]; however, adapting them to handle multiple snapshots is non-trivial.

**Dense subgraphs of evolving graphs:** Existing work also studies the problem of identifying a dense subgraph on time-evolving graphs [3, 4, 7]; these are graphs where new nodes and edges may appear over time and existing ones may disappear. The goal in this line of work is to devise a *streaming algorithm* that at any point in time it reports the densest subgraph for the current version of the graph. In our work, we are not interested in the dynamic version of the problem and thus the algorithmic challenges that our problem raises are orthogonal to those faced by the work on streaming algorithms.

Table 8: The hashtags output as solutions to the  $O^2$ BFF problem on *Twitter*.

<b>k = 3</b>	BFF-MM, BFF-MA	BFF-AM	BFF-AA
	kimi, abudhabigp, fl, allowin	ozpol, nz, mexico, malaysia, singapore, vietnam, chile, peru, tpp, japan, canada	abudhabigp, fp1, abudhabi, guti, fl, pushpush, skyfl, hulk, allowin, bottas, kimi, fp3, fp2
<i>Dates:</i>	<i>Oct 31-Nov 2</i>	<i>Oct 27-28, Nov 7</i>	<i>Oct 31-Nov 2</i>
<b>k = 6</b>	BFF-MM, BFF-MA	BFF-AM	BFF-AA
	abudhabigp, fl, skyfl	wikileaks, snowden, nsa, prism	abudhabigp, fp1, abudhabi, guti, fl, pushpush, skyfl, hulk, allowin, bottas, kimi, fp3, fp2
<i>Dates:</i>	<i>Oct 28-Nov 2</i>	<i>Oct 27-28, Nov 3,5,7</i>	<i>Oct 28, Oct 30-Nov 1, Nov 9</i>
<b>k = 9</b>	BFF-MM, BFF-MA	BFF-AM	BFF-AA
	(Too many tags to report)	wikileaks, snowden, nsa, prism	assange, wikileaks, snowden, nsa, prism
<i>Dates:</i>		<i>Oct 27-31, Nov 3,5-7</i>	<i>Oct 27-29,31, Nov 3,5-7,10</i>
<b>k = 12</b>	BFF-MM, BFF-MA	BFF-AM	BFF-AA
	(Too many tags to report)	wikileaks, snowden, nsa	assange, wikileaks, snowden, nsa, prism
<i>Dates:</i>		<i>Oct 27-Nov 1, Nov 3-7,10</i>	<i>Oct 27-31, Nov 2-7, 10</i>

## 8. SUMMARY

In this paper, we introduced and systematically studied the problem of identifying dense subgraphs in a collection of graph snapshots defining a graph history. We showed that for many definitions of aggregate density functions the problem of identifying a subset of nodes that are densely-connected in *all* snapshots (i.e., the BFF problem) can be solved in linear time. We also demonstrated that other versions of the BFF problem (i.e., BFF-MA and BFF-AM) cannot be solved with the same algorithm. To identify dense subgraphs that occur in  $k$ , yet not all, the snapshots of a history graph we also defined the  $O^2$ BFF problem. For all variants of this problem we showed that they are NP-hard and we devised an iterative algorithm for solving them. Finally, we studied extensions of these problems that incorporate query nodes and connectivity constraints. Our extensive experimental evaluation with datasets from diverse domains (e.g., collaboration networks, social and computer networks) demonstrated the effectiveness and the efficiency of our algorithms.

## 9. REFERENCES

- [1] J. I. Alvarez-Hamelin, L. Dall’Asta, A. Barrat, and A. Vespignani. Large scale networks fingerprinting and visualization using the k-core decomposition. In *Advances in Neural Information Processing Systems, NIPS*, pages 41–50, 2005.
- [2] Y. Asahiro, K. Iwama, H. Tamaki, and T. Tokuyama. Greedily finding a dense subgraph. In *Scandinavian Workshop on Algorithm Theory, SWAT*, pages 136–148, 1996.
- [3] B. Bahmani, R. Kumar, and S. Vassilvitskii. Densest subgraph in streaming and mapreduce. *PVLDB*, 5(5):454–465, 2012.
- [4] S. Bhattacharya, M. Henzinger, D. Nanongkai, and C. E. Tsourakakis. Space- and time-efficient algorithm for maintaining dense subgraphs on one-pass dynamic streams. In *Annual ACM on Symposium on Theory of Computing, STOC*, pages 173–182, 2015.
- [5] J. Bourjolly, G. Laporte, and G. Pesant. An exact algorithm for the maximum k-club problem in an undirected graph. *European Journal of Operational Research*, 138(1):21–28, 2002.
- [6] M. Charikar. Greedy approximation algorithms for finding dense components in a graph. In *Approximation Algorithms for Combinatorial Optimization*, pages 84–95, 2000.
- [7] A. Epasto, S. Lattanzi, and M. Sozio. Efficient densest subgraph computation in evolving graphs. In *WWW*, pages 300–310, 2015.
- [8] A. V. Goldberg. *Finding a maximum density subgraph*. University of California Berkeley, CA, 1984.
- [9] V. Jethava and N. Beerenwinkel. Finding dense subgraphs in relational graphs. In *ECML PKDD*, pages 641–654, 2015.
- [10] S. Khuller and B. Saha. On finding dense subgraphs. In *International Colloquium on Automata, Languages and Programming, ICALP*, pages 597–608, 2009.
- [11] J. Leskovec, J. M. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *TKDD*, 1(1), 2007.
- [12] K. Makino and T. Uno. New algorithms for enumerating all maximal cliques. In *Scandinavian Workshop on Algorithm Theory, SWAT*, pages 260–272, 2004.
- [13] B. McClosky and I. V. Hicks. Combinatorial algorithms for the maximum k-plex problem. *J. Comb. Optim.*, 23(1):29–49, 2012.
- [14] P. Rozenstein, N. Tatti, and A. Gionis. Discovering dynamic communities in interaction networks. In *ECML PKDD*, pages 678–693, 2014.
- [15] M. Sozio and A. Gionis. The community-search problem and how to plan a successful cocktail party. In *ACM SIGKDD*, pages 939–948, 2010.
- [16] P. Tsantarlitis and E. Pitoura. Topic detection using a critical term graph on news-related tweets. In *Proceedings of the Workshops of the EDBT/ICDT*, pages 177–182, 2015.
- [17] C. E. Tsourakakis, F. Bonchi, A. Gionis, F. Gullo, and M. A. Tsiarli. Denser than the densest subgraph: extracting optimal quasi-cliques with quality guarantees. In *ACM SIGKDD*, pages 104–112, 2013.

## APPENDIX

In this section we present counter-examples that demonstrate that the  $\text{FINDBFF}_M$  and  $\text{FINDBFF}_A$  when applied to the BFF-MA and BFF-AA yield a solution that is a poor approximation of the optimal solution. For the following, we use  $n = |V|$  to denote the number of nodes in the different snapshots, and  $\tau = |\mathcal{G}|$  to denote the number of snapshots.

### A. PROOF OF PROPOSITION 4

*Proof.* In order to prove our claim we need to construct an instance of the BFF-AM problem where the  $\text{FINDBFF}_M$  algorithm produces a solution with approximation ratio  $O(\frac{1}{n})$ . We construct the graph history  $\mathcal{G} = \{G_1, \dots, G_\tau\}$  as follows. The first  $\tau - 1$  snapshots consist of a full clique with  $n - 1$  nodes, plus an additional node  $v$  that is connected to a single node  $u$  from the clique. The last snapshot  $G_\tau$  consists of just the edge  $(v, u)$ .

In the first  $n - 2$  iterations of the  $\text{FINDBFF}_M$  algorithm, the node with the minimum degree is one of the nodes in the clique (other than the node  $u$ ). Thus the nodes in the clique will be iteratively removed, until we are left with the edge  $(u, v)$ . Since node  $v$  is present in all intermediate subsets  $S_i$ , the minimum degree in all snapshots  $G_t$  is 1. Therefore, the solution  $S$  of the  $\text{FINDBFF}_M$  algorithm has  $f_{am}(S) = 1$ . On the other hand clearly the optimal solution  $S^*$  consists of the nodes in the clique, where we have minimum degree  $n - 2$ , except of the last instance where the minimum degree is zero. Therefore,  $f(S^*) = (n - 2) \frac{\tau - 1}{\tau}$  which proves our claim.  $\square$

### B. PROOF OF PROPOSITION 5

*Proof.* In order to prove our claim we need to construct an instance of the BFF-AM problem where the  $\text{FINDBFF}_A$  algorithm produces a solution with approximation ratio  $O(\frac{1}{n})$ . We construct the graph history  $\mathcal{G} = \{G_1, \dots, G_\tau\}$ , where  $\tau$  is even, as follows. Each snapshot  $G_t$  contains  $n = 2b + 3$  nodes. The  $2b$  of these nodes form a complete  $b \times b$  bipartite graph. Let  $u, v$ , and  $s$  denote the additional three nodes. Node  $s$  is connected to all nodes in the graph, in all snapshots, except for the last snapshot where  $s$  is connected only to  $u$  and  $v$ . Nodes  $u$  and  $v$  are connected to each other in all snapshots, and node  $u$  is connected to all  $2b$  nodes of the bipartite graph in the first  $\tau/2$  snapshots, while node  $v$  is connected to all  $2b$  nodes of the bipartite graph in the last  $\tau/2$  snapshots. Throughout assume that  $\tau \geq 2$ . Note that the optimal set  $S^*$  for this history graph consists of the  $2b$  nodes in the bipartite graph, with  $f_{am}(S^*, \mathcal{G}) = b = \Theta(n)$ .

The score  $score_a$  for every node  $w$  of the  $2b$  nodes in the bipartite graph is  $score_a(w, \mathcal{G}) = b + 1 + \frac{\tau - 1}{\tau}$ . For the nodes  $u$  and  $v$ , we have  $score_a(u, \mathcal{G}) = score_a(v, \mathcal{G}) = \frac{2b\tau/2 + 2\tau}{\tau} = b + 2$ . Node  $s$  has score  $score_a(s, \mathcal{G}) = 2b \frac{\tau - 1}{\tau} + 2$ .

Therefore, in the first iteration, the algorithm will remove one of the nodes of the bipartite graph. Without loss of generality assume that it removes one of the nodes in the left partition. Now, for a node  $w$  in the left partition, we still have that  $score_a(w, \mathcal{G}[S_1]) = b + 1 + \frac{\tau - 1}{\tau}$ . For a node  $w$  in the right partition we have that  $score_a(w, \mathcal{G}[S_1]) = b + \frac{\tau - 1}{\tau}$ . For nodes  $u$  and  $v$  we have  $score_a(u, \mathcal{G}[S_1]) = score_a(v, \mathcal{G}[S_1]) = \frac{(2b - 1)\tau/2 + 2\tau}{\tau} = b + \frac{3}{2}$ . For node  $s$  we have that  $score_a(s, \mathcal{G}[S_1]) = (2b - 1) \frac{\tau - 1}{\tau} + 2$ .

Therefore, in the second iteration the algorithm will select to remove one of the nodes in the right partition. Note that the resulting graph  $\mathcal{G}[S_2]$  is identical in structure with  $\mathcal{G}$ , with  $n = 2(b - 1) + 3$  nodes. Therefore, the same procedure will be repeated until all the nodes from the bipartite graph are removed, while nodes  $u$  and  $v$  will be kept in the set until the last iterations. As a result, the set  $S$  returned by  $\text{FINDBFF}_A$  has  $f_{am}(S, \mathcal{G}) = 2$  (the degree of the nodes  $u$  and  $v$ ), yielding approximation ratio  $O(\frac{1}{n})$ .  $\square$

### C. PROOF OF PROPOSITION 6

*Proof.* In order to prove our claim we need to construct an instance of the BFF-MA problem where the  $\text{FINDBFF}_M$  algorithm produces a solution with approximation ratio  $O(\frac{1}{\sqrt{n}})$ . We construct the graph history  $\mathcal{G} = \{G_1, \dots, G_\tau\}$  as follows. We have  $\tau = m$  snapshots that are all identical. They consist of two sets of nodes  $A$  and  $B$  of size  $m$  and  $m^2$  respectively. The nodes in  $B$  form a cycle. The nodes in  $A$  in graph snapshot  $G_t$  form a clique with all nodes except for one node  $v_t$ , different for each snapshot. The optimal set  $S^*$  consists of the nodes in  $A$ , that have average degree  $\frac{(m - 1)(m - 2)}{m} = \Theta(m)$ .

The  $\text{FINDBFF}_M$  starts with the set of all nodes. The average degree of any snapshot is  $\frac{2m^2 + (m - 1)(m - 2)}{m^2 + m} = \Theta(1)$ , which is also the value of the  $f_{ma}(V)$  function. In the first  $m$  iterations of the algorithm, the nodes in  $A$  have  $score_m(v, S_i) = 0$ , so these are the ones to be removed first. Then the nodes in  $B$  are removed. In all iterations the average degree in each snapshot remains  $O(1)$ . Therefore, the set  $S$  returned by the  $\text{FINDBFF}_M$  has  $f_{ma}(S) = \Theta(1)$ , and the approximation ratio is  $\Theta(\frac{1}{m})$ . Since  $m = \sqrt{n}$ , this proves our claim.  $\square$

### D. PROOF OF PROPOSITION 7

*Proof.* In order to prove our claim we need to construct an instance of the BFF-MA problem where the  $\text{FINDBFF}_A$  algorithm produces a solution with approximation ratio  $O(\frac{1}{\sqrt{n}})$ . The construction of the proof is very similar to before. We construct the graph history  $\mathcal{G} = \{G_1, \dots, G_\tau\}$  as follows. We have  $\tau = m$  snapshots that are all identical, except for the last snapshot  $G_m$ . The snapshots  $G_1, \dots, G_{m-1}$  consist of two sets of nodes  $A$  and  $B$  that form two complete cliques of size  $m$  and  $m^2$  respectively. In the last snapshot the nodes in  $B$  are all disconnected. The optimal set  $S^*$  consists of the nodes in  $A$ , that have  $f_{ma}(A) = \frac{m(m - 1)}{m} = \Theta(m - 1)$ .

The  $\text{FINDBFF}_A$  starts with the set of all nodes. The value of  $f_{ma}(V)$  is determined by the last snapshot  $G_m$  that has average degree  $\frac{m(m - 1)}{m^2 + m} = \Theta(1)$ . The nodes in  $A$  have average degree (over time)  $\frac{m(m - 1)}{m} = \Theta(m)$ , while the nodes in  $B$  have average degree  $\frac{(m - 1)(m^2 - 1)}{m} = \Theta(m^2)$ . Therefore, the algorithm will iteratively remove all nodes in  $A$ . In each iteration the resulting set  $S_i$  has  $f_{ma}(S_i) = O(1)$ . When all the nodes in  $A$  are removed, we have that  $f_{ma}(S_i) = 0$ . Therefore, the approximation ratio for this instance is  $\Theta(\frac{1}{m})$ . Our claim follows from the fact that  $n = m^2 + m$ .  $\square$